Predicting Continuous Conflict Perception with Bayesian Gaussian Processes

Samuel Kim, Fabio Valente, Maurizio Filippone *Member, IEEE* and Alessandro Vinciarelli *Member, IEEE*

Abstract—Conflict is one of the most important phenomena of social life, but it is still largely neglected by the computing community. This work proposes an approach that detects common conversational social signals (loudness, overlapping speech, etc.) and predicts the conflict level perceived by human observers in continuous, non-categorical terms. The proposed regression approach is fully Bayesian and it adopts Automatic Relevance Determination to identify the social signals that influence most the outcome of the prediction. The experiments are performed over the SSPNet Conflict Corpus, a publicly available collection of 1430 clips extracted from televised political debates (roughly 12 hours of material for 138 subjects in total). The results show that it is possible to achieve a correlation close to 0.8 between actual and predicted conflict perception.

Index Terms—Social Signal Processing, Conflict, Gaussian Processes, Automatic Relevance Determination.

1 Introduction

Whenever it takes place, interpersonal conflict influences the life of groups to a significant extent, most often with negative consequences [1]. In the workplace, conflict spans from minor disagreements to physical assault and, in all cases, is one of the most important causes of stress [2]. At home, marital conflict is a major source of distress and, if not properly handled, can lead families to disintegration [3]. In general, interpersonal conflict is likely to cause long-term, negative effects on the rapport between individuals [4].

For socially intelligent technologies, expected to understand and seamlessly integrate human interactions [5], [6], predicting conflict perception can be the first step towards dealing appropriately with the phenomenon. In particular, domains that can benefit from conflict measurement are, e.g., automatic analysis of content in multimedia data [7], meeting analysis [8], social robotics [9] and any other area where automatic understanding of human-human interactions can play a role.

This work focuses on televised political debates. The *rationale* behind such a choice is that debates are often built around conflict (e.g., between two or more competing candidates) and the chances of observing the phenomenon, possibly including a dose of incivility [10], have been increasing during the last years [11]. In this respect, political debates might be similar to other situations (e.g., work meetings) where social norms are tight and impose behavioral limitations, but issues at stake are important and conflict is still observable.

The literature shows that individuals involved in conflict tend to display both positive and negative emotions as well as different levels of arousal [12]. Furthermore, the main effect of emotions in conflict seems to be the choice of strategies (e.g., avoidance vs. engagement) that can correspond to different behavioral displays [13], [14]. Therefore, this work privileges social signals that, following the literature, tend to be less ambiguous as conflict markers [3], [15], [16], [17], [18].

According to different theoretic orientations, social signals correspond to "*acts or structures that influence the behavior or internal state of other individuals*" [19], "*communicative or informative signals which [...] provide information about social facts*" [20], or "*actions whose function is to bring about some reaction or to engage in some process*" [21]. In other words, social signals are *observable* behaviors that not only convey information about social phenomena, but also influence others and their behaviors.

Following the Social Signal Processing methodology [5], this work proposes an approach that automatically detects social signals typical of conversations and, based on their frequency and intensity, it predicts the conflict level perceived in the social interactions under analysis. The social signals most likely to account for the conflict level were identified with crowdsourcing techniques (551 annotators involved via Mechanical Turk) and then represented with features automatically extracted from the data. The literature shows that automatic speech transcriptions do not result in deteriorated emotion recognition even when the word error rate is significant [22]. This is likely to apply to automatic conflict perception as well, but focusing on the sole nonverbal communication is a common approach and can still lead to good results (see [6] for an extensive survey).

The experiments were performed over the *SSPNet Conflict Corpus*, a publicly available collection of 1430

- *S.Kim and Fabio Valente are with Idiap Research Institute. E-mail: {skim,fvalente}@idiap.ch*
- *M.Filippone is with University of of Glasgow. E-mail: Maurizio.Filippone@glasgow.ac.uk*
- *A.Vinciarelli is with University of Glasgow and Idiap Research Institute. E-mail: Alessandro.Vinciarelli@glasgow.ac.uk*
- *This work was supported by the European Commission through the Network of Excellence SSPNet (www.sspnet.eu).*

clips extracted from a database of political debates (roughly 12 hours of material including 138 subjects). The results show a correlation close to 0.8 between automatically predicted and manually annotated conflict level.

To the best of our knowledge, this is the first work that not only deals with conflict in *dimensional* terms, but also it proposes a Bayesian approach for Automatic Relevance Determination (ARD) in Gaussian Processes [23], i.e. for weighting the features according to their influence on the regression output. This is an improvement with respect to previous work on the *SSPNet Conflict Corpus* that was based on categorical approaches [7] or regression techniques without ARD [24]. In the proposed approach, features are first pruned out of the model by adopting Maximum Likelihood optimization; second, full characterization of the posterior distribution of the ARD parameters is carried out based on Markov chain Monte Carlo (MCMC). This is motivated by the difficulties in characterizing the full posterior distribution of such parameters [25], especially in the application considered here, which involves a large number of ARD parameters.

The rest of this paper is organized as follows: Section 2 proposes a survey of previous work in the literature, Section 3 describes the data collection and annotation process, Section 4 presents the automatic detection of social signals in speech, Section 5 describes the approach for the conflict level prediction, Section 6 reports on experiment and results and the final Section 7 draws some conclusions.

2 Previous Work

The computing community is making significant efforts towards the development of socially intelligent machines that sense and understand the social landscape like humans do [5]. The literature proposes a large number of approaches dealing with some of the most important social and psychological phenomena (see [6] for an extensive survey), but conflict has received only limited attention because it is difficult to access ecologically valid data [26]. For this reason, earlier works focused on *agreement* and *disagreement*, easier to observe and annotate, while actual conflict detection and measurement approaches appeared only recently.

2.1 Disagreement Detection

Agreement and disagreement are defined as a relation of congruence or opposition, respectively, between opinions expressed by multiple parties involved in the same interaction [27]. The detection of disagreement (see [28] for a survey) is relevant to conflict analysis because the two phenomena, while being different, often co-occur. Most of the experiments were performed over meeting recordings [29], [30], [31], [32], [33], but recent work shifted towards political debates, a scenario where conflict and disagreement are more likely to take place [34]. Table 1 provides a synopsis of the main works available in the literature.

The approach proposed in [29] adopts heuristic features accounting for both verbal and non-verbal aspects of interaction. The former include number and type ("*positive*" and "*negative*") of words, as well as the perplexity of statistical language models trained over both agreement and disagreement samples. The latter include fundamental frequency statistics (maximum, minimum and average) and duration of speech "*spurts*" ("*a period of speech by one speaker that has no pauses of greater than one half second*" according to the definition given in the work). The classification is performed using decision trees and the accuracy (percentage of correctly classified spurts) is 61%. The approach proposed in [32] uses the same data as in [29], but it adopts a Maximum-Entropy ranking technique for the classification of spurts. The features include speaker adjacency statistics (e.g., number of spurts between interventions of two speakers), duration modeling (e.g., amount of time a speaker talks) and lexical measurements (e.g., the number of words in a spurt). The accuracy in a four-way classification task (including disagreement among classes) is 84.0%.

The experiments of [30], [31] aim at automatic detection and classification of "*hot spots*", meeting segments where participants are particularly engaged (including disagreement moments). The first work [30] uses dialogue acts, word counts and perplexity of language models trained over large corpora of written text as features. The detection of disagreement hot spots is then performed with decision trees and the chance normalized accuracy goes up to 0.4. The second work [31] identifies deviations of fundamental frequency and energy as a reliable evidence of several hot spots, including disagreement. In the same vein, the features of [33] include dialogue acts (not only of the segment to be classified, but also of the neighboring ones to take into account the context), lexical choices (e.g. part of speech tags and key-words selected via an effectiveness ratio) and prosody (energy, pitch and speech rate). Agreement and disagreement detection, performed using decision trees and Conditional Random Fields, leads to an $F1$ measure close to 45%. The F1 measure corresponds to $2\alpha\beta/(\alpha+\beta)$, where α is the precision (probability that a sample assigned to a class actually belongs to that class) and β is the recall (probability that a sample actually belonging to a class is assigned to that class).

An attempt to go beyond the simple classification of agreement and disagreement episodes was proposed in [34], where Hidden State Conditional Random Fields are applied to investigate the dynamics of disagreement in political debates. The input cues are prosody (energy and pitch) as well as automatically detected gestures. The maximum accuracy achieved is close to 65%.

2.2 Conflict Detection

One of the reasons why early work focused on disagreement is that meeting scenarios are often co-operative

Ref.	Subjects	Behavioral Cues	Phenomenon	Annotation	Data	Performance
[7]	138	Turn Organization Prosody Speaker Adjacency Stats.	conflict	categorical	SSPNet Conflict Corpus	$F1 = 76.1\%$ clip accuracy (3 classes)
[24]	138	Turn Organization Prosody Speaker Adjacency Stats.	conflict	dimensional	SSPNet Conflict Corpus	correlation 0.75 predicted / real conflict level
[29]	40-50	Prosody Lexical	(dis)agreement	categorical	9854 spurts ICSI Meetings	61% accuracy
[30]	53	Dialogue Acts Lexical	hot spots	categorical	32 ICSI meetings	0.4 chance normalized accuracy
[31]	20-30	Prosody	hot spots	categorical	13 ICSI meetings	significant correlation
[32]	40-50	Duration, Lexical Speaker Adjacency	(dis)agreement	categorical	9854 spurts ICSI Meetings	84% accuracy
[33]	16	Prosody, Lexical Dialogue Acts	(dis)agreement	categorical	20 AMI Meetings	$F1 \sim 45\%$
[34]	44	Prosody Gestures	(dis)agreement	categorical	147 Debate clips from Canal9	64.2% accuracy
[36]	26	Turn Organization Steady Conversational Periods	conflict	categorical	13 Debates from Canal9	80.0% turn classification accuracy
[37]	138	Overlapping Speech to Non-Overlapping Speech Ratio	conflict	categorical	SSPNet Conflict Corpus	$UAR = 83.1\%$ clip accuracy (2 classes)
[38] (1)	138	Feature Selection Over OpenSmile Acoustic Features	conflict	categorical	SSPNet Conflict Corpus	$UAR = 83.9\%$ clip accuracy (2 classes)
[38] (2)	138	Feature Selection Over OpenSmile Acoustic Features	conflict	dimensional	SSPNet Conflict Corpus	correlation 0.82 predicted / real conflict level
[39]	26	Lexical	blaming acceptance	categorical	130 Couple Therapy Sessions	$> 70.0\%$ classification accuracy

TABLE 1
The table shows the most important works dedicated to conflict and disagreement. The performances are reported for the sake of completeness, but they cannot be compared because they are not always obtained over the same data.

(like in the case of the AMI meetings) or involve individuals unlikely to engage in conflict [33]. Hence, it is not surprising to observe that approaches explicitly aimed at the detection of conflict appeared only recently, when data portraying conflictual interactions became available [28], [35]. In particular, recent work has turned towards political debates [7], [24], [36], [37], [38] and couple therapy sessions [39], two settings where conflict, defined as an interaction process where different parties pursue incompatible goals (see Section 3 for more details), is a more frequent phenomenon. Table 1 provides details of the main works available in the literature.

Experiments and approaches presented in [7], [36] deal with categorical definitions of conflict. In the case of [36], conflict is considered present or absent, while the other work considers three possible levels (absent-to-low, middle, high). The approach is based on "*Steady Conversational Periods*", i.e. statistical representations of stable conversational configurations (e.g., everybody talks, one person talks and the others listen, etc.). An approach based on Generative Score Spaces [40] allows the authors to segment the data into "*conflict / non-conflict intervals*". The percentage of data time correctly labeled in such terms is 80%.

The work in [7] uses the same data of this work (see Section 3.1), but it adopts a categorical approach. The paper takes into account prosodic features (statistics from pitch, energy and articulation rate), speaker adjacency statistics, overlapping speech and turn-organization. Then, it applies Support Vector Machines to assign clips extracted from political debates to one of the three classes mentioned above. The resulting $F1$ score is 76.1%. The work in [24] uses the same features as [7], but it adopts a dimensional representation of conflict. Therefore, the goal of the work is not the classification or the detection, but the measurement of conflict level. A regression approach based on Gaussian Processes allows the authors to reach a correlation close to 0.8 between actual and predicted conflict level.

Two conflict related tasks, based on the data of this work (see Section 3.1), were proposed at the "*Interspeech 2013 Computational Paralinguistics Challenge*" [41]: the binary classification of the samples into *high* and *low* conflict and the prediction of the continuous conflict level associated to each sample. The classification task was addressed in [37], [38]. In the first work [37], the experiments show that Unweighted Average Recall

(UAR) values higher than 80% can be achieved by using only one feature, namely the ratio of overlapping speech to non-overlapping speech (the value of the feature is predicted using 6373 acoustic features provided by the challenge organizers [41]). In the second work [38], the application of a random subset selection approach to the 6373 acoustic features above leads to a UAR of 83.9%. The same feature selection approach is used to predict the continuous conflict level as well and the resulting correlation between actual and predicted value is 0.82.

The last approach [39] works on a large corpus of couple therapy sessions and predicts the attitude of one spouse towards the other as perceived by observers. It adopts lexical features (frequency of appearance of words used by each subject) to identify, among others, blaming or acceptance attitudes, possibly accounting for the presence or absence of conflict, respectively. Accuracies higher than 70% are achieved for both classes.

3 Conflict and its Perception

The definitions proposed in the literature are multiple and diverse, but they tend to agree on one point, namely that conflict takes place whenever multiple parties involved in an interaction pursue incompatible goals (or at least perceive this to happen): "*conflict is a process in which one party perceives that its interests are being opposed or negatively affected by another party*" [42], "*[conflict takes place] to the extent that the attainment of the goal by one party precludes its attainment by the other*" [43], "*Conflict is perceived [...] as the perceived incompatibilities by parties of the views, wishes, and desires that each holds*" [13], etc.

Goals, interests, views, etc. are not accessible to observation, but they influence behavior. Therefore conflict can be perceived and detected, at least in principle, through its effect on the way people behave, including social signals being displayed [3], [15], [16], [17], [18]. For this reason, the annotation process applied in this work aims at "measuring" the link between observable, possibly machine detectable social signals, and the level of conflict as perceived by human observers.

3.1 The Data

The experiments of this work are performed over televised political debates (see Section 1 for the motivations). Television material "*can engender the neglect of minimal requirements for experimental control of important determinants*" [44]. However, it can be considered a reliable alternative to field data [45] and it is often used for research on emotions [46] or nonverbal behavioral cues like, e.g., facial expressions [47]. Furthermore, debate participants are likely to have incompatible goals: if one politician gets elected, the other does not, if one party acquires consensus, the other loses it, etc. Therefore, according to the definition provided at the beginning of Section 3, the probability of observing conflict in the data should be sufficiently high.

In particular, the data used in this work were extracted from "*Canal9*", a database of political debates televised in Switzerland during 2005 [35]. The Canal9 debates were segmented into uniform, non-overlapping windows of 30 seconds and only the segments portraying at least two persons were retained. Compared to shorter windows or analysis units, 30 seconds long segments are less ambiguous and, therefore, the annotations are more likely to converge. The result is a collection of 1430 clips - the *SSPNet Conflict Corpus* - showing 138 subjects for a total length of 11 hours and 55 minutes. The data is publicly available[1] and it was used as a benchmark for the "*Interspeech 2013 Computational Paralinguistics Challenge*" (see Section 2) [41].

The length of the clips is an empirical tradeoff between two conflicting needs: the first is that the windows must be long enough to have a reasonable chance of including at least two speakers (otherwise it is not possible to observe conflict), the second is that the windows must be short enough to cover, at least partially, only one conflict episode. Given that the average turn-length (a turn is a time interval during which only one person speaks) in the Canal9 Corpus is 19.7 seconds, the use of 30 seconds long segments appears to address both needs to a reasonable extent. An indirect confirmation comes from the experiments of [37], where the clips of the SSPNet Conflict Corpus were split into three 10 seconds long segments to analyze patterns of escalation and de-escalation: the three windows of each clip were labeled as *High* (H) or *Low* (L) in terms of conflict. The pattern $\{\text{HLH}\}$, the only one that can account for two conflict episodes (one in the first 10 seconds and the other one in the last 10 seconds) was observed only 4.4% of the times (63 clips out of 1430).

3.2 The Annotation Questionnaire

In this work, the goal of the annotation is to measure how social signals influence conflict perception in human observers. For this reason, the annotation questionnaire adopted in the experiments consists of two *layers*: the first one, called *physical*, includes questions about observable, detectable and measurable conflict markers (see below). The second, called *inferential*, includes questions about the interpretation of a scene in terms of competition and conflict. The questions are listed in Table 2, in the same order as when they were administered during the annotation process; each item is associated to a 5-points Likert scale mapped into the interval $[-2, 2]$.

The questions of the physical layer take into account the social signals that the literature shows to be frequently associated to conflict. Items Q2, Q9 and Q13 consider interruptions and overlapping speech, typically used to grab, hold and possibly steal the floor [17], [18]. Questions Q3 and Q6 assess fast speaking and loudness that typically accompany conflictual interactions [15], [16]. Questions Q4, Q7 and Q10 consider the overall level

1. http://sspnet.eu/2013/09/sspnet-conflict-corpus/

#	Question	Layer
Q1	The atmosphere is relaxed (-)	I
Q2	People wait for their turn before speaking (-)	P
Q3	One or more people talk fast (+)	P
Q4	One or more people fidget (+)	P
Q5	People argue (+)	I
Q6	One or more people raise their voice (+)	P
Q7	One or more people shake their heads and nod (+)	P
Q8	People show mutual respect (-)	I
Q9	People interrupt one another (+)	P
Q10	One or more people gesture with their hands (+)	P
Q11	One or more people are aggressive (+)	I
Q12	The ambience is tense (+)	I
Q13	One or more people compete to talk (+)	P
Q14	People are actively engaged (+)	I
Q15	One or more people frown (+)	P

TABLE 2
The table shows the questionnaire used to annotate the clips of the corpus. The first column reports the question ID, the second column shows the question with its sign and the third column says whether the question belongs to the Inferential (I) or Physical (P) layer.

of motor activity that tends to increase in the presence of conflict [3], [15]. Finally, question Q15 addresses the use of facial expressions conveying negative affect [3]. The questions of the inferential layer aim at measuring the intensity of the conflict as perceived by the annotators. If the cues considered in the physical layer actually account for the presence of conflict, physical and inferential scores - the sum of the answers given to items in the physical and inferential layer, respectively - should show significant correlation (see next section).

3.3 Crowdsourcing Annotation

The data annotation was performed via Amazon Mechanical Turk (AMT), one of the most commonly adopted crowdsourcing platforms. The 1430 clips of the SSPNet Conflict Corpus were randomly split into 143 groups of 10 samples each. The 143 groups were used to create 143 HITs[2] (*Human Intelligence Tasks*), the individual tasks that an annotator must perform to receive a payment. In this case, every HIT, i.e. the annotation of a group of 10 clips, was rewarded with 1 US Dollar.

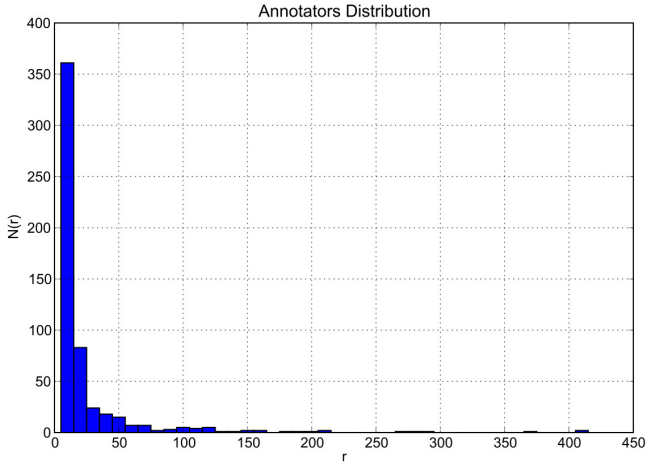

Fig. 1. The plot shows the number $N(r)$ of annotators that have judged r clips.

The inclusion of 10 clips in a HIT aims at detecting non-cooperative annotators, i.e. those that fill the questionnaire of Table 2 randomly. Two pairs of items, {Q1,Q12} and {Q2,Q9}, are repetitions of the same statement in opposite terms (e.g., "*The atmosphere is relaxed*" and "*The ambience is tense*"). Therefore, the sum of the answers to such items over a HIT should be close to zero. If such a condition is not met and the sum is significantly different from zero, the annotator is likely to be non-cooperative. Thanks to this control mechanism, around 20% of the submitted questionnaires were discarded.

The HITs were assigned randomly to the annotators and they were removed once they were performed by 10 cooperative annotators. In this way, each clip of the SSPNet corpus has been assessed 10 times. No limitation was imposed to the number of HITs that annotators were allowed to perform. However, most of these latter performed only one HIT (361 out of the 551) and Figure 1 shows the resulting distribution of the number of clips per annotator (only annotators retained after applying the control mechanism above are taken into account).

Since the work focuses on nonverbal communication, it is necessary to limit as much as possible the effect of what people say in the clips of the Corpus. To this purpose, research on nonverbal communication adopts different methods. In some cases, speech recordings are split into short frames (e.g., 10 ms) that are then locally re-shuffled to make words non-understandable while preserving nonverbal vocal behavior [48]. In other cases, the subjects are asked to utter meaningless sequences of syllables like if they were real words [49].

This work adopts the approach proposed in [50], where assessors are asked to annotate material in a language they do not understand. In particular, the clips of the *SSPNet Conflict Corpus* are in French, but only US annotators were allowed to work on the data. Before their first HIT, the annotators were asked to state whether they understood French or not. In case of positive answer, an annotator was not allowed to perform any of the 143 HITs. It was not possible to check whether the annotators answered honestly or not, but the last available report of the US census bureau states that only 0.5% of the US population, roughly 1.6 millions of people, speaks French [3]. Annotating Swiss data in the US might introduce a cross-cultural bias, i.e. a systematic disalignment between the way American and Swiss observers judge the same situation [51]. Furthermore, other effects, difficult to predict, cannot be excluded.

At the end of the annotation process, there are 10 filled

2. https://www.mturk.com/mturk/welcome
3. http://www.census.gov/prod/2003pubs/c2kbr-29.pdf

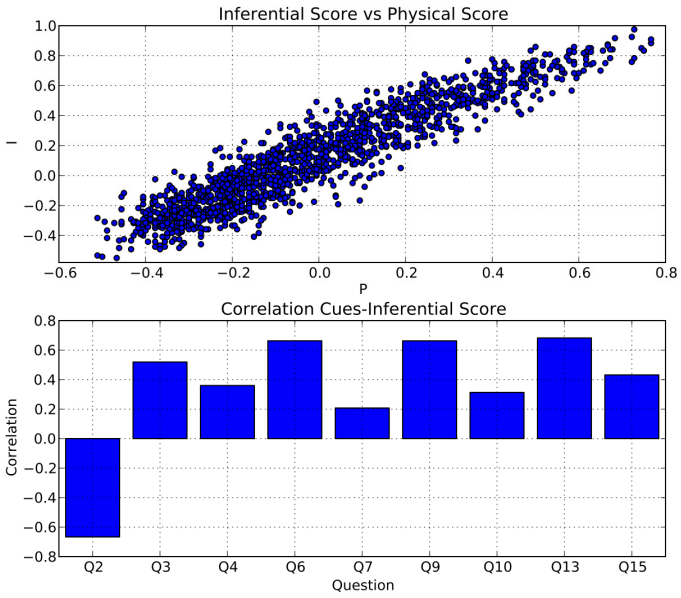

Fig. 2. The upper plot shows the correlation between physical and inferential scores for each clip. The lower plot shows the correlation between each question of the physical layer and the inferential score.

questionnaires per clip (Figure 1 shows the distribution of the number of clips per assessor). Each questionnaire provides two scores: the *physical* one is the sum of the answers to the questions of the physical layer, the *inferential* one is the sum of the answers to the questions of the inferential layer. Since there are 10 questionnaires per clip, there are 10 physical and 10 inferential scores as well. The average of the former corresponds to the overall physical score of the clip, the average of the latter corresponds to its inferential score. In Figure 2, each point of the upper plot corresponds to one of the 1430 clips and the coordinates are the overall physical and inferential scores mentioned above. The correlation between the two scores is 0.95 ($p = 10^{-12}$) and this suggests that the cues assessed in the physical layer tend to co-occur with the assignment of a high inferential score. This latter plays the role of continuous label in the prediction experiments (see Section 6).

The agreement between assessors was measured in terms of *effective reliability* R [52]:

$$R = \frac{Nr}{1+(N-1)r} \quad ; \quad r = 2\frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} r_{ij}}{N(N-1)} \qquad (1)$$

where N is the number of assessors and r is the average of the correlations between all possible pairs of assessors (r_{ij} is the correlation between assessors i and j). The value of R is 0.91 and 0.92 for physical and inferential score, respectively (the average correlation between assessors is 0.52 in the first case and 0.53 in the second).

4 Feature Extraction

The annotation of the data (see Section 3) does not provide only the conflict level observed in each sample of the corpus, but also an indication on the social signals most likely to influence the perception of the scenes. The lower plot of Figure 2 shows the correlation between inferential score and scores obtained for each question of the physical layer individually (all values are statistically significant with $p < 1\%$ according to a t-test). The lowest absolute values correspond to items Q4 ("*One or more people fidget*"), Q7 ("*One or more people shake their heads and nod*"), Q10 ("*One or more people gesture with their hands*") and Q15 ("*One or more people frown*"), i.e. the cues that can be detected in the video channel. Therefore, the feature extraction process will focus on the audio channel while not considering the video one. The most probable reason for the difference between speech and other cues is that televised data allow observers to listen to everybody (the microphones are always open for all participants), but not to see everybody (the camera shows only what the director decides to show). Therefore, observers might be induced to rely on what they hear more than on what they see.

The feature extraction process includes two main stages. The first is the *speaker diarization* and aims at segmenting audio recordings into *turns*, i.e. intervals where only one person speaks. This step is necessary to extract features that account for turn-organization (*who talks when, how much and with whom*) and behavior of individual debate participants. The second is the actual feature extraction step and aims at representing the clips of the corpus with a vector of measurements accounting for the cues investigated in Section 3.

In this work, the diarization is performed with an agglomerative clustering approach based on the Information Bottleneck principle (see [53] for a full description). The process is unsupervised and it is not necessary to know in advance the number of speakers talking in the data. During the diarization, each acoustic observation is assigned to one speaker only (corresponding to one of the clusters). Therefore, the detection of overlapping speech segments - where at least two speakers talk at the same time - requires a further processing step aimed at detecting audio segments including multiple voices (see [54] for a full description). The experiments are performed both with and without overlapping speech detection to estimate the effect of such a cue (see Section 6).

The diarization performance is measured in terms of purity π, a segmentation effectiveness metric showing, on one hand, to what extent all feature vectors corresponding to a given category are assigned the same label in the segmentation and, on the other hand, to what extent all feature vectors assigned the same label in the segmentation actually belong to the same category. In the experiments of this work, categories correspond to speakers or overlapping speech and labels are arbitrary

identifiers. The purity was adopted because it allows one to consider all overlapping speech segments equivalent, i.e. to avoid the identification of the speakers that talk at the same time. The value of π ranges between 0 and 1 (the higher the better) and it is the geometric mean of two terms: the *average cluster purity* π_c and the *average speaker purity* π_s. The definitions of π_c and π_s are as follows:

$$\pi_c = \sum_{k=1}^{N_c} \sum_{l=1}^{N_s} \frac{n_k}{N} \frac{n_{lk}^2}{n_k^2} \quad ; \quad \pi_s = \sum_{l=1}^{N_s} \sum_{k=1}^{N_c} \frac{n_l}{N} \frac{n_{lk}^2}{n_l^2} \tag{2}$$

where N is the total number of feature vectors, N_s is the number of categories (all speakers and overlapping speech), N_c is the number of clusters (each corresponding to a label) detected in the diarization process, n_{lk} is the number of vectors belonging to category l that have been attributed to cluster k, and n_k is the number of feature vectors in cluster k. In the experiments of this work, $\pi = 0.8$ before the application of the overlapping speech detector and $\pi = 0.82$ after.

Once the diarization is complete, it is possible to extract the actual features that account for prosody (90 features), turn-duration statistics (10 features), speaker adjacency statistics (5 features) and, when the detection algorithm is applied, overlapping speech (5 features).

4.1 Prosodic Features

Short term prosodic features, in particular pitch (measured with the algorithm proposed in [55]) and intensity, are extracted with Praat[4] from $30\ ms$ long frames at regular time steps of $10\ ms$. This results into 3×10^3 measurements per clip that are then represented through their statistical properties.

Clip-based pitch and intensity statistics (18 features): they include pitch and intensity mean, median, standard deviation, minimum, maximum and quantiles (0.01, 0.25, 0.75 and 0.99) computed over the entire clip. Before computing clip statistics, frame-level prosodic features are speaker normalized by applying the *Z-norm*: $\bar{x} = (x - m_s)/\sigma_s$ where m_s and σ_s are speaker mean and standard deviation obtained on the entire debate from which the clip is extracted. These features are expected to capture loudness and speaking styles typically accompanying conflictual interactions [15], [16]. These features account in particular for item Q6 in the questionnaire (see Section 3).

Turn-based pitch and intensity statistics (54 features): they include the same nine statistics as above, but applied to mean, median and standard deviation of pitch and intensity extracted turn-by-turn. These features account for the same behavioural aspects mentioned at the previous point, but aim at capturing long-term aspects during the clip. In this case as well, the features correspond to item Q6 of the questionnaire.

4. http://www.praat.org/

Overlapping speech pitch and intensity statistics (18 features): The nine statistics above (mean, median standard deviation, minimum, maximum and quantiles) are applied to pitch and intensity extracted from overlapping speech segments. These features are expected to account for speaking behaviour during overlapping speech, one of the most salient aspects of competitive discussions [17], [18]. These features correspond to items Q2, Q9 and Q13 in the questionnaire.

4.2 Conversational Features

After the diarization, the clips are segmented into turns and overlapping speech segments. This makes it possible to extract features that account for turn-organization:

Turn duration statistics (6 features): they include number of turns, mean, median, standard deviation, minimum and maximum of turn durations over the clip. Turn duration can provide information about the tendency to talk for shorter intervals of time during conflicts or competitive discussions [18]. These features correspond to items Q2 and Q13 of the questionnaire.

Speaking duration statistics (6 features): they include number of speakers in the clip, mean, median, standard deviation, minimum and maximum of the total speaking time of each individual in the clip. These features will provide further information about the overall regime of the conversation [56]. These features correspond to items Q2, Q3 and Q13 of the questionnaire.

Speaker adjacency statistics (3 features): each participant in the discussion is either the moderator (m), or a participant belonging to one of the two groups (g_1 and g_2) opposing one another in the debate. The bigram probabilities $p(r_t|r_{t-1})$ where $r_i \in \{m, g_1, g_2\}$ is the "*role*" of the speaker at turn i are used to build the following features: $p(m|g_1) + p(m|g_2)$, the probability of the moderator grabbing the floor, $p(g_1|m) + p(g_2|m)$, the probability of one of the participants grabbing the floor after the moderator, and $p(g_1|g_2) + p(g_2|g_1)$ probability of an exchange between participants. These statistics aim at capturing preference structures related to conflict and, in particular, the tendency to react immediately to others we disagree with [57]. However, these features are available only when using a manual - and not automatic - diarization (see Section 6 for more details). These features correspond to items Q9 and Q13 in the questionnaire.

Overlapping speech duration statistics (4 features): they include the fractions of the clip corresponding to overlapping speech, overlapping speech involving moderator and participants, overlapping speech involving members of the same group (see above) and members of different groups. The amount of overlapping speech is important because it tends to increase when there is competition to grab and hold the floor like it happens in conflicts [17], [18]. These features correspond to items Q2, Q9 and Q13 in the questionnaire.

Turn keeping/turn stealing ratio (1 feature): The ratio between the number of times that the speaker is the same before and after an interval of overlapping speech and the number of times that, after an overlapping speech interval, the speaker changes. This measure accounts for how frequently debate participants try to dominate the conversation and prevent others from expressing their opinions [58]. This feature corresponds to item Q9 in the questionnaire.

5 Gaussian Process Regression

At the end of the feature extraction process, each clip i is represented by a feature vector $\mathbf{x}_i$. Given that the conflict level y_i is a continuous variable (it corresponds to the inferential score described in Section 3.3), the mapping between $\mathbf{x}_i$ and y_i can be inferred with a regression approach. This work focuses in particular on Gaussian Processes (GPs) with Automatic Relevance Determination [23], [59]. The reason is that such models allow one to perform nonparametric nonlinear regression and to identify the features that influence most the mapping between features and corresponding target observations. Hence, it is possible to investigate, at least indirectly, the nonverbal cues that influence most the conflict level prediction. In the rest of this section, $\boldsymbol{\theta}$ denotes model parameters and $\mathbf{y} = \{y_1, \ldots, y_n\}$ is the vector comprising the conflict level y_i for each clip.

Typically, regression approaches model y as a random variable with mean given by the sum of a latent (unobserved) function f of $\mathbf{x}$, parametrized by $\boldsymbol{\theta}$, and a stochastic term ε, resulting in $y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon$, where ε is usually assumed Gaussian distributed. The main characteristic of a regression model is the way the latent function $f(\mathbf{x}, \boldsymbol{\theta})$ is specified. In parametric models, latent functions are constructed as a combination of basis functions. Unless there are reasons to believe that a set of basis functions is adequate in explaining the mapping between features and labels, a nonparametric specification of $f(\mathbf{x}, \boldsymbol{\theta})$ is more appealing. In the application considered here, for example, it is not clear how to determine a set of basis functions capable of modeling the relationship between features and conflict level. We therefore propose to employ GPs as they allow for a nonparametric modeling of $f(\mathbf{x}, \boldsymbol{\theta})$. Formally, a GP is a set of random variables characterized by the property that any finite subset of them is jointly distributed as a Gaussian, and specifying its mean and covariance functions is enough to completely characterize it. Following the most common approach in the literature, this work models latent variables using a zero mean GP prior. The choice of the covariance function influences the properties of the functions that can be modeled by GPs, such as smoothness and range of values that the functions span, as discussed next.

5.1 Covariance functions

To give an intuition on the meaning of the covariance function, imagine one that rapidly decays with the distance between input locations; this leads to functions that can rapidly change between nearby input values, due to the low covariance between the function at these locations. In a regression setting, when ε is distributed as a Gaussian, the GP assumption on $f(\mathbf{x}, \boldsymbol{\theta})$ implies that $\mathbf{y}$ can be modelled directly as a multivariate Gaussian $\mathcal{N}(\mathbf{y}|\mathbf{0}, K)$, where K is a $n \times n$ covariance matrix evaluated at the input vectors.

In the experiments, two covariance functions are tested that yield a nonlinear mapping between features vectors $\mathbf{x}_i$ and conflict level y_i. Both can be expressed in a Radial Basis Function (RBF) form:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_a \exp\left[-(\mathbf{x}_i - \mathbf{x}_j)^T \Lambda (\mathbf{x}_i - \mathbf{x}_j)\right] + \delta_{ij}\theta_{\sigma^2}$$

where δ_{ij} returns 1 when $i = j$ and zero otherwise. We will refer to the spherical case $\Lambda = \theta_{\text{global}} I$ as the RBF covariance. The second type of covariance that we will consider, uses a diagonal matrix $\Lambda = \text{diag}(\boldsymbol{\theta}_{\text{ARD}})$ and yields the RBF with Automatic Relevance Determination (ARD) [59], [23]; we will refer to this as the RBF ARD covariance. RBF and RBF ARD covariance functions decay with distance at a rate that depends on the choice of the parameters in Λ. The RBF function has only one global parameter controlling the decay of the covariance function, while the RBF ARD function has one parameter for each feature. The main advantage of this latter solution is that the values of the different parameters account for the influence of the features on the predicted value y: the larger the parameter the higher the influence of the corresponding feature. In this respect, the RBF ARD covariance can provide indications on the nonverbal cues that most influence the perception of conflict, a property particularly desirable for Social Signal Processing applications.

5.2 Predictions and inference of parameters

For simplicity of notation, let $\boldsymbol{\theta}$ be the set of all parameters parameterizing $k(\mathbf{x}_i, \mathbf{x}_j)$. Consider a test feature vector $\mathbf{x}_*$, and define the covariance matrix K, the vector $\mathbf{k}_*$ whose ith element is $k(\mathbf{x}_i, \mathbf{x}_*)$, namely the covariance between the test and the ith feature vector, and $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$. Under the GP modelling assumptions, the label y_* associated to $\mathbf{x}_*$ is distributed as a Gaussian with mean $\mathbf{k}_*^{\mathrm{T}} K^{-1} \mathbf{y}$ and variance $k_{**} - \mathbf{k}_*^{\mathrm{T}} K^{-1} \mathbf{k}_*$.

Before any predictions are made, given a training set it is necessary to adapt the model parameters $\boldsymbol{\theta}$. Usually, this is carried out by optimization of the log-likelihood $\log[p(\mathbf{y}|X, \boldsymbol{\theta})]$ with respect to $\boldsymbol{\theta}$, where X denotes the set of all $\mathbf{x}_i$. Optimizing the parameters, however, can lead to underestimation of the uncertainty in predictions, and in a wrong assessment of the relative importance of the different features [23], [60], [61], so we propose to adopt a fully probabilistic approach able to overcome these limitations. In order to do so, the following integral needs to be solved:

$$p(y_*|\mathbf{y}, X, \mathbf{x}_*) = \int p(y_*|\mathbf{x}_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}, X) d\boldsymbol{\theta}, \qquad (3)$$

which requires the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y}, X)$ encoding the uncertainty in model parameters; this allows for a sound quantification of uncertainty in the assessment of the importance of the different features as shown in the results. For GP regression (GPR) it is not feasible to carry out this computation analytically and it is necessary to resort to some approximation.

We propose to draw N samples from $p(\boldsymbol{\theta}|\mathbf{y}, X)$ denoted by $\boldsymbol{\theta}^{(i)}$ and to use the Monte Carlo approximation

$$p(y_*|\mathbf{y}, X, \mathbf{x}_*) \simeq \frac{1}{N}\sum_{i=1}^{N} p(y_*|\mathbf{x}_*, \boldsymbol{\theta}^{(i)}). \quad (4)$$

Such an approximation yields the desired integral in the limit of N going to infinity, which in practice gives the possibility to achieve results to the desired level of precision given N large enough. As it is not possible to draw samples from $p(\boldsymbol{\theta}|\mathbf{y}, X)$ directly, we employ Markov chain Monte Carlo (MCMC) methods, and in particular the standard Metropolis-Hastings (MH) algorithm (see, e.g., [62] for full details).

6 Experiments and Results

The experiments were performed over the data presented in Section 3 using the features described in Section 4. In particular, three different variants of the feature extraction process were considered: The first, called "*Manual*", extracts prosodic and conversational features after manually segmenting the data into turns and overlapping speech segments. The second, called "*Automatic*", extracts the same features, but after applying an automatic speaker diarization process that does not distinguish between turns and overlapping speech. The third, called "*Automatic w.o.s.*" ("*w.o.s.*" stands for "*with overlapping speech*"), extracts the features after applying not only the speaker diarization, but also an overlapping speech detector.

For the sake of comparison, the conflict level prediction was performed not only with the GP based approach described above (with and without ARD), but also with two other approaches, namely Bayesian Linear Regression (BLR) [23] and Support Vector Regression (SVR) [63]. We used the SVR approach with the standard ε-insensitive loss function [63] as implemented in the `LIBSVM` library [64]. In the experiments, parameters of BLR and SVR were tuned using cross-validation within the training set. The fact that optimization of kernel parameters in SVR is carried out by means of cross-validation makes it unfeasible to employ the ARD kernel.

The performance was measured in terms of correlation between actual and predicted conflict perception as well as Root Mean Square Error (see Section 6.1). Furthermore, the coefficients of the ARD covariance were used to identify the features most likely to influence the prediction of the GP regression and, indirectly, the cues most likely to influence human observers perception.

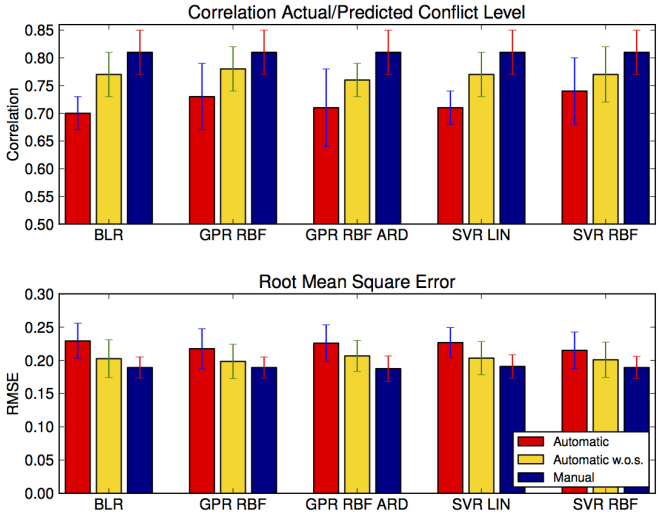

Fig. 3. Correlation coefficients (upper part) and Root Mean Square Errors (lower part) achieved with different regression approaches for manual diarization, automatic diarization, and automatic diarization with overlapping speech. The error bars correspond to the standard deviation computed across the five folds.

6.1 Performance metrics

Let m be the number of test samples, and let $\hat{y}_i$ represent the prediction for the ith test point with actual target value y_i. Also, let $\hat{\mu}$ and μ be the mean values of $\hat{y}_i$ and y_i across the test set, and $\hat{\sigma}^2$ and σ^2 the variances of $\hat{y}_i$ and y_i across the test set. The two evaluation metrics used to assess the performance in predicting the conflict level are the Correlation Coefficient (CC):

$$\mathrm{CC} = \frac{1}{m\,\sigma\,\hat{\sigma}}\sum_{i=1}^{m}(y_i - \mu)(\hat{y}_i - \hat{\mu}) \quad (5)$$

and the Root Mean Square error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}. \quad (6)$$

6.2 Results

Figure 3 reports the results obtained using the regression methods and feature extraction processes described in the previous sections. The performance of the tested regression approaches was measured by employing 5-fold cross-validation. The samples were distributed across the 5 folds by ensuring speaker and debate independence, i.e. by avoiding that the same subject and/or the same debate appear in both training and test set. For each fold, model parameters were optimized by maximizing the cross-validation error on the training set alone; note that in contrast to BLR and SVR approaches, the GPR approach has no regularization parameters to optimize.

All regression models achieve the best performance with the *Manual* feature extraction process. The reason

is that in such a setting there are fewer errors in the detection of speaker changes and overlapping speech segments. Therefore, prosodic features are extracted from speech segments that actually correspond to one voice (or to overlapping speech) and conversational features correspond to the actual statistics observed in the data. Furthermore, it is possible to assign a role to the speakers (see Section 4) and this seems to contribute to the correct prediction of the conflict level (see Section 6.3).

The performance loss in going from the Manual feature extraction process to the Automatic w.o.s. one is only significant (p-value < 0.05 in a paired t-test) in the case of GPR with the RBF ARD covariance. When going from the Automatic w.o.s. extraction process to the Automatic one, the loss in performance is always significant (p-value < 0.05 in a paired t-test) except in the case of SVR with the RBF kernel. This confirms the indications of Figure 2, where the questions related to interruptions and competition for speaking ("*People wait for their turn before speaking*", "*People interrupt one another*" and "*One or more people compete to talk*"), probably the main sources of overlapping speech, show absolute values of correlation higher than 0.6 with the inferential score. Overall, RMSE and correlation between actual and predicted inferential score are around 0.2 and 0.75, respectively. The RMSE obtained when predicting always the average observed inferential score is 0.35. The regression approaches used in the experiments have similar performance, but the one proposed in this work (GPR ARD) has the important advantage of showing the features that have the highest impact on the regressor outcome. The next section shows how this can help to better understand the interplay between nonverbal cues and conflict perception.

6.3 Interpretation of the ARD coefficients

One of the main aspects of Bayesian Learning is that model parameters are treated as random variables and they are inferred from data (see Section 5.2). In the case of the GP approach with Automatic Relevance Determination, this means that a full probability distribution on the parameters weighting each feature in the covariance matrix K is estimated. In the experiments of this work, this allows one to estimate how each feature and, indirectly, each nonverbal cue influences the predicted conflict level. Furthermore, this assessment is carried out in a probabilistic way, rather than by optimization. This is of fundamental importance to avoid misinterpretations on the role played by the features, as in optimization one would only draw conclusions on the one configuration of the parameters yielding the optimal fit to the data.

The analysis of the ARD coefficients has several advantages over the analysis of correlations, the technique typically adopted to measure the association between features and ratings. Correlation is linear (hence unable to reflect more complex relationships), sensitive to outliers (one sample may be sufficient to change significantly its value) and can be applied only to individual features (it cannot take into account an entire set of variables like the method proposed here). Furthermore, ARD coefficients are directly related to the functioning of the regression approach and show the features most likely to improve the correlation between actual and predicted conflict perception. In a technology oriented experiment, this is a particularly desirable property, especially because it is possible to do so while employing a nonparametric nonlinear regression approach.

For the data of this article, the number of ARD covariance parameters is 110 (one for each feature). This makes it difficult to apply MCMC not only for the high dimensionality of the feature vectors, but also because the weights of less relevant features will be sharply peaked around zero, an obstacle towards the efficient exploration of the parameter space. Therefore, the approach proposed in this work includes two steps. The first identifies non-relevant ($\theta < 0.1$) features with a Maximum-Likelihood approach. The second carries out fully probabilistic inference of the remaining covariance parameters by sampling from their posterior distribution using the MH algorithm.

The priors imposed on the parameters, Gamma functions $\mathrm{Ga}(\theta_r|1,1)$, were weakly informative. The sampling was applied on a log-transformed version of the parameters to avoid dealing with boundary conditions (e.g., positivity). According to common practices in MCMC, convergence to the posterior distribution was assessed by analyzing the $\hat{R}$ potential scale reduction factor [65] based on 10 parallel chains. Running the chains for 25000 iterations with a burn-in phase of 5000 iterations (where chains were allowed to adapt and reach around 25% acceptance rate) was sufficient to reach convergence.

The results appear in Figure 4, where the boxplots show the posterior distribution of the parameters obtained at the second stage of the training process above. The vertical line of the box corresponds to the median of the posterior distribution over each parameter and the whiskers extend from the lower to the upper quartile. The higher the median, the higher the influence of the corresponding feature on the predicted conflict level. The boxplots are visible only for those parameters that were not discarded after the first optimization stage, namely $\theta > 0.1$. The parameters are grouped according to the meaning of the features they weight (see right side of the plots and Section 4).

In the case of *Manual*, the most important feature seems to be the minimum of the intensity (turn-based), meaning that clips where people speak louder (higher intensity minimum) tend to be perceived as more conflictual. Similar considerations apply to the minima of pitch (both turn- and clip-based) and turn-based intensity. Minima are likely to be affected by noise due to errors in pitch and intensity estimate, but the indications seem to confirm not only the literature on nonverbal correlates of conflict [15], [16], but also the indications of Figure 2 showing that question Q6 ("*One or more people raise their voice*") on loudness is one of the most

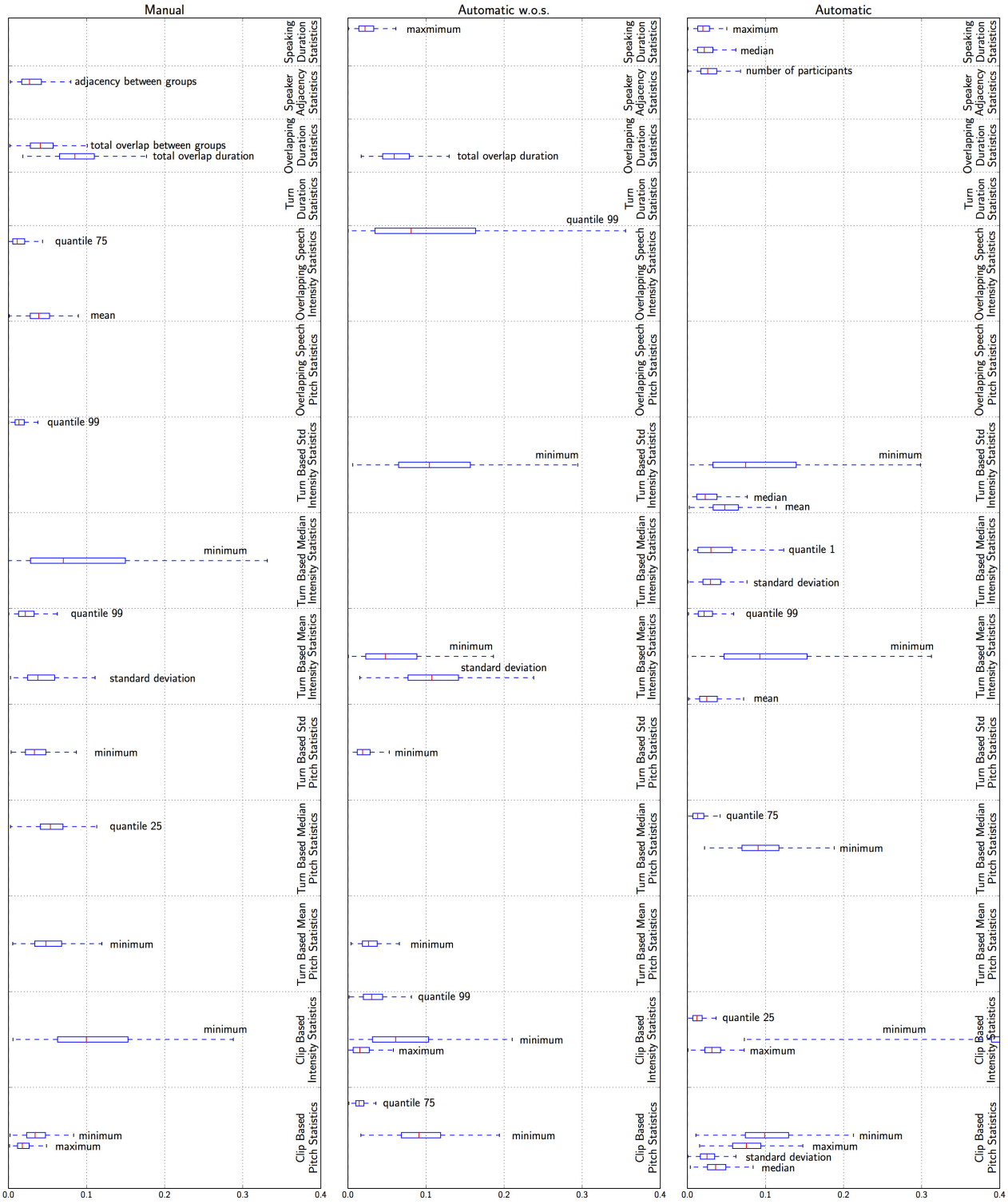

Fig. 4. The boxplots account for the distribution of the θ_i parameters of the RBF ARD covariance matrices. The red vertical line is the median of the distribution and the whiskers range between the lower and upper quartile. The higher the median, the higher the influence of the corresponding feature. Each of the three plots includes 110 parameters, but only those with median significantly different from zero are shown. The right hand side of the plot shows how features are grouped according to the description of Section 4.

correlated with the inferential score.

The role of the intensity minima might account for *entrainment*, the *"speaker's adaptation to the speech of his interlocutor"* [66]. In other words, it might happen that

the participants converge to the intensity of the loudest speakers and, as a result, the minimum of the intensity increases. On the other hand, the features of this work do not consider the speakers individually and, therefore, it is not possible to conclude whether the minima tend to be high because people match their respective intensities (as expected in the case of the entrainment) or because there is an escalation where speakers try to be louder than each other. Furthermore, previous results show that entrainment tends to be more frequent in interactions where the mutual attitude of the participants is positive [67] and, during conflicts, this is typically not the case. However, conclusions can be made only by taking into account the intensities of the individual spakers and not, like in this work, global statistics across all of them.

The total duration of overlapping speech, especially if it involves members of different groups in the debate (see description of speaker adjacency statistics in Section 4), confirms that overlapping speech is one of the most salient markers of conflict [17], [18]. Furthermore, it is in line with the results of [37], where the ratio of overlapping speech to non-overlapping speech is sufficient to discriminate clips with high and low level of conflict (see Section 2). Last, but not least the probability of finding speakers belonging to different groups one after another in the speaker sequence indicates the actual presence of preference structures [57] such that individuals tend to react to others they disagree with more than to others they agree with. In turn, observing such a preference structure elicits the perception of higher conflict levels. This appears in the speaker adjacency statistics part of Figure 4 (upper part of the "*Manual*" plot).

The observations about the minimum of pitch and intensity made for *Manual* apply to *Automatic* as well. In particular, the minimum of the intensity over the clip plays such an important role that the median is outside the range of the plot (all plots have the same range for the sake of clarity). The inevitable errors in the diarization process determine more noise (see in particular the large number of relevant features among the clip-based statistics). However, the minima tend to have a higher median. Since no overlapping speech detector is applied in *Automatic*, features related to such a phenomenon do not have any influence. In contrast, some features that did not appear to be relevant in the *Manual* case seem to have an influence here. In particular, median and maximum of the speaking time for each subject suggest that during conflictual interactions more people tend to talk for a longer time, probably as a result of the competition for grabbing the floor typically observed in conflicts [18]. The total number of participants seems to have an influence as well, but it is probably a spurious effect due to the errors of the diarization process. In fact, when there are more interruptions or overlapping speech, the clustering algorithm behind the diarization tends to find more clusters that are interpreted as more voices and, then, more speakers.

If the diarization process is followed by an overlapping speech detection (*Automatic w.o.s.*), the considerations made so far about the minima of pitch and intensity do not change, but the 99% quantile of the intensity during overlapping speech segments (an approximation of the maximum) plays for the first time a role. This shows that it is not sufficient to talk together to convey the impression of an on-going conflict, but it is also necessary to speak louder. In this case as well, both psychological literature and indications of the crowdsourcing annotation are confirmed [17].

7 Conclusions

Conflict is one of the most important social phenomena [1] and this article proposes an approach for the measurement of its perception during face-to-face interactions. The experiments were performed over the *SSPNet Conflict Corpus* and the results show that the correlation between actual and predicted conflict level is between 0.7 and 0.8 (see upper plot in Figure 3), corresponding to a Root Mean Square Error of roughly 0.2 (see lower plot in Figure 3).

To the best of our knowledge, this is the first time that conflict is defined in dimensional rather than categorical terms. This appears to be particularly suitable for this problem because the conflict levels observed in the data are distributed continuously and do not cluster around two or more modes possibly corresponding to different classes. In this respect, the data annotation methodology presented in this work - inspired by established behavior observation techniques - allows one to deal more effectively with naturalistic situations where conflict takes time to (de)-escalate and does not simply switch on and off. Furthermore, it allows one to take into account situations where the intensity of the conflict changes according to the importance of the issues being debated.

Besides the regression performance, the application of Automatic Relevance Determination made it possible to identify the features with the highest influence on the model outcome. The results show that the model predictions are in agreement not only with the observations done during the annotation of the data, but also with the literature on nonverbal correlates of conflict.

Given the importance of conflict in everyday life [2], [3], [4], the development of approaches capable of sensing the phenomenon can be of interest for socially intelligent technologies expected to sense the interaction landscape and react appropriately to it [5], [6]. Improvements of the approach proposed in this work might come from three main directions. The first is the inclusion of cues extracted from the video channel (e.g., facial expressions or gestures), the second is the refinement of the inference approaches and the third is the extraction of better features from the data. Furthermore, this work focused on nonverbal behavior, but useful information can certainly come from the analysis of the verbal content of the interactions.

REFERENCES

[1] J. M. Levine and R. L. Moreland, "Small groups," in *The handbook of social psychology*, D. Gilbert and G. Lindzey, Eds. Oxford University Press, 1998, vol. 2, pp. 415–469.

[2] P. Spector and S. Jex, "Development of four self-report measures of job stressors and strain: interpersonal conflict at work scale, organizational constraints scale, quantitative workload inventory, and physical symptoms inventory," *Journal of Occupational Health Psychology*, vol. 3, no. 4, pp. 356–367, 1998.

[3] J. Gottman, H. Markman, and C. Notarius, "The topography of marital conflict: A sequential analysis of verbal and nonverbal behavior," *Journal Of Marriage And The Family*, vol. 39, no. 3, pp. 461–477, 1977.

[4] R. Baumeister, A. Stillwell, and S. Wotman, "Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger," *Journal of Personality and Social Psychology*, vol. 59, no. 5, pp. 994–1005, 1990.

[5] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image and Vision Computing Journal*, vol. 27, no. 12, pp. 1743–1759, 2009.

[6] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud, I. Poggi, F. D'Errico, and M. Schroeder, "Bridging the Gap Between Social Animal and Unsocial Machine: A Survey of Social Signal Processing," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 69–87, 2012.

[7] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 5089–5092.

[8] S. Renals, H. Bourlard, J. Carletta, and A. Popescu-Belis, Eds., *Multimodal Signal Processing: Human Interactions in Meetings*. Cambridge University Press, 2012.

[9] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and autonomous systems*, vol. 42, no. 3, pp. 143–166, 2003.

[10] D. Mutz and B. Reeves, "The new videomalaise: Effects of televised incivility on political trust," *American Political Science Review*, vol. 99, no. 1, pp. 1–15, 2005.

[11] C. Reinemann and M. Maurer, "Unifying or polarizing? short-term effects and postdebate consequences of different rhetorical strategies in televised debates," *Journal of Communication*, vol. 55, no. 4, pp. 775–794, 2005.

[12] W. Arsenio and M. Killen, "Conflict-Related Emotions During Peer Disputes," *Early Education and Development*, vol. 7, no. 1, pp. 43–57, 1996.

[13] C. Bell and F. Song, "Emotions in the Conflict Process: an Application of the Cognitive Appraisal Model of Emotions to Conflict Management," *International Journal of Conflict Management*, vol. 16, no. 1, pp. 30–54, 2005.

[14] P. Jordan and A. Torth, "Managing Emotions During Team Problem Solving: Emotional Intelligence and Conflict Resolution," *Human Performance*, vol. 17, no. 2, pp. 195–218, 2004.

[15] A. L. Sillars, S. F. Coletti, D. Parry, and M. A. Rogers, "Coding Verbal Conflict Tactics: Nonverbal and Perceptual Correlates of the "Avoidance-Distributive-Integrative" Distinction," *Human Communication Research*, vol. 9, no. 1, pp. 83–95, 1982.

[16] V. Cooper, "Participant and observer attribution of affect in interpersonal conflict: an examination of noncontent verbal behavior," *Journal of Nonverbal Behavior*, vol. 10, no. 2, pp. 134–144, 1986.

[17] L. Smith-Lovin and C. Brody, "Interruptions in Group Discussions: The Effects of Gender and Group Composition," *American Sociological Review*, vol. 54, no. 3, pp. 424–435, 1989.

[18] E. Schegloff, "Overlapping Talk and the Organisation of Turn-taking for Conversation," *Language in Society*, vol. 29, no. 1, pp. 1–63, 2000.

[19] M. Mehu and K. Scherer, "A psycho-ethological approach to Social Signal Processing," *Cognitive Processing*, vol. 13, no. 2, pp. 397–414, 2012.

[20] I. Poggi and F. D'Errico, "Social Signals: a framework in terms of goals and beliefs," *Cognitive Processing*, vol. 13, no. 2, pp. 427–445, 2012.

[21] P. Brunet and R. Cowie, "Towards a conceptual framework of research on Social Signal Processing," *Journal of Multimodal User Interfaces*, vol. 6, no. 3-4, pp. 101–115, 2012.

[22] F. Metze, A. Batliner, F. Eyben, T. Polzehl, B. Schuller, and S. Steidl, "Emotion recognition using imperfect speech recognition." in *Proceedings of Interspeech*, 2010, pp. 478–481.

[23] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[24] S. Kim, M. Filippone, F. Valente, and A. Vinciarelli, "Predicting the conflict level in television political debates: an approach based on crowdsourcing, nonverbal communication and Gaussian processes," in *Proceedings of the ACM International Conference on Multimedia*, 2012, pp. 793–796.

[25] A. Vehtari and J. Lampinen, "Bayesian input variable selection using posterior probabilities and expected utilities," *Report B31*, 2002.

[26] A. Vinciarelli, S. Kim, F. Valente, and H. Salamin, "Collecting Data for Socially Intelligent Surveillance and Monitoring Approaches: The Case of Conflict in Competitive Conversations," in *Proceedings of International Symposium on Communications, Control and Signal Processing*, 2012, pp. 1–4.

[27] I. Poggi, F. D'Errico, and L. Vincze, "Agreement and its Multimodal Communication in Debates: A Qualitative Analysis," *Cognitive Computation*, vol. 3, no. 3, pp. 466–479, 2011.

[28] K. Bousmalis, M. Mehu, and M. Pantic, "Towards the Automatic Detection of Spontaneous Agreement and Disagreement Based on Nonverbal Behaviour: A Survey of Related Cues, Databases and Tools," *Image and Vision Computing Journal (to appear)*, 2012.

[29] D. Hillard, M. Ostendorf, and E. Shriberg, "Detection of Agreement vs. Disagreement in Meetings: Training with Unlabeled Data," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2003.

[30] B. Wrede and E. Shriberg, "The Relationship Between Dialogue Acts and Hot Spots in Meetings." in *Proceedings of the IEEE Speech Recognition and Understanding Workshop*, 2003, pp. 180–185.

[31] ——, "Spotting"Hotspots" in Meetings: Human Judgments and Prosodic Cues," in *Proceedings of Eurospeech*, 2003, pp. 2805–2808.

[32] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg, "Identifying Agreement and Disagreement in Conversational Speech: Use of Bayesian Networks to Model Pragmatic Dependencies," in *Proceedings of the Meeting of the Association for Computational Linguistics*, 2004.

[33] S. Germesin and T. Wilson, "Agreement Detection in Multiparty Conversation," in *Proceedings of ACM International Conference on Multimodal Interfaces*, 2009, pp. 7–14.

[34] K. Bousmalis, L. P. Morency, and M. Pantic, "Modeling Hidden Dynamics of Multimodal Cues for Spontaneous Agreement and Disagreement Recognition," in *Proceedings of Face and Gesture*, 2011, pp. 746–752.

[35] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: a Database of Political Debates for Analysis of Social Interactions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, vol. 2, 2009, pp. 96–99.

[36] A. Pesarin, M. Cristani, V. Murino, and A. Vinciarelli, "Conversation analysis at work: Detection of conflict in competitive discussions through automatic turn-organization analysis," *Cognitive Processing*, vol. 13, no. 2, pp. 533–540, 2012.

[37] F. Grezes, J. Richards, and A. Rosenberg, "Let me finish: Automatic conflict detection using speaker overlap," in *Proceedings of Interspeech*, 2013.

[38] O. Räsänen and J. Pohjalainen, "Random subset feature selection in automatic recognition of developmental disorders, affective states, and level of conflict from speech," in *Proceedings of Interspeech*, 2013.

[39] P. Georgiou, M. Black, A. Lammert, B. Baucom, and S. Narayanan, ""That's Aggravating, Very Aggravating": Is It Possible to Classify Behaviors in Couple Interactions Using Automatically Derived Lexical Features?" in *Proceedings of International Conference on Affective Computing and Intelligent Interaction*, 2011, pp. 87–96.

[40] A. Perina, M. Cristani, U. Castellani, V. Murino, and N. Jojic, "Free energy score space," in *Advances in Neural Information Processing Systems*, 2009, pp. 1428–1436.

[41] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M.Mortillaro, H.Salamin, A.Polychroniou, F.Valente, and S.Kim, "The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings of Interspeech*, 2013.

[42] J. Wall and R. Roberts Callister, "Conflict and its management," *Journal of Management*, vol. 21, no. 3, pp. 515–558, 1995.
[43] C. Judd, "Cognitive Effects of Attitude Conflict Resolution," *Journal of Conflict Resolution*, vol. 22, no. 3, pp. 483–498, 1978.
[44] K. Scherer, "Vocal markers of emotion: Comparing induction and acting elicitation," *Computer Speech and Language*, vol. 27, no. 1, pp. 40–58, 2013.
[45] P. Waxer, "Video ethology: television as a data base for cross-cultural studies in nonverbal displays," *Journal of Nonverbal Behavior*, vol. 9, no. 2, pp. 111–120, 1985.
[46] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1, pp. 227–256, 2003.
[47] P. Stewart, F. Salter, and M. Mehu, "Taking leaders at face value: Ethology and the analysis of televised leader displays," *Politics and the Life Sciences*, vol. 28, no. 1, pp. 48–74, 2009.
[48] K. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *European Journal of Social Psychology*, vol. 8, no. 4, pp. 467–487, 1978.
[49] T. Bänziger and K. Scherer, "Introducing the Geneva Multimodal Emotion Portrayal (GEMEP) corpus," in *Blueprint for affective computing: A sourcebook*, K. Scherer, T. Bänziger, and E. Roesch, Eds. Oxford University Press Oxford, 2010, pp. 271–294.
[50] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–278, 2012.
[51] D. Lehman, C. Chiu, and M. Schaller, "Psychology and culture," *Annual Review of Psychology*, pp. 689–714, 2004.
[52] R. Rosenthal, "Conducting judgment studies: Some methodological issues," in *The new handbook of methods in nonverbal behavior research*, J. Harrigan, R. Rosenthal, and K. Scherer, Eds. Oxford University Press, 2005, pp. 199–234.
[53] D. Vijayasenan, F. Valente, and H. Bourlard, "An Information Theoretic Approach to Speaker Diarization of Meeting Data," *IEEE Transactions on Audio Speech and Language Processing*, vol. 17, no. 7, pp. 1382–1393, 2009.
[54] K. Boakye, O. Vinyals, and G. Friedland, "Improved Overlapped Speech Handling for Speaker Diarization," in *Proceedings of Interspeech*, 2011, pp. 941–944.
[55] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proceedings of the Institute of Phonetic Sciences*, vol. 17, no. 1193, 1993, pp. 97–110.
[56] G. Yule, *Pragmatics*. Oxford University Press, 1996.
[57] J. Bilmes, "The concept of preference in conversation analysis," *Language in Society*, vol. 17, pp. 161–181, 1988.
[58] M. Rahim, "A measure of styles of handling interpersonal conflict," *Academy of Management Journal*, vol. 26, no. 2, pp. 368–376, 1983.
[59] C. E. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.
[60] M. Filippone, A. Marquand, C. Blain, S. Williams, J. Mourão-Miranda, and M. Girolami, "Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities," *Annals of Applied Statistics*, vol. 6, no. 4, pp. 1883–1905, 2012.
[61] M. Filippone and M. Girolami, "Exact-approximate Bayesian inference for Gaussian processes," 2013, arXiv:1310.0740.
[62] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer-Verlag, 2005.
[63] V. Vapnik, *The nature of statistical learning theory*. Springer-Verlag, 1995.
[64] C. Chang and C. Lin, "LIBSVM: A library for Support Vector Machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, 2011.
[65] A. Gelman and D. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Science*, vol. 7, no. 4, pp. 457–472, 1992.
[66] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions." in *Proceedings of Interspeech*, 2011, pp. 3081–3084.
[67] C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples." in *Proceedings of Interspeech*, 2010, pp. 793–796.

Samuel Kim received his Ph.D. on Electrical Engineering from University of Southern California, Los Angeles, California in 2010 focusing on contextual modeling of audio signals toward information retrieval. He received his B.S. and M.S. on Electrical and Electronic Engineering from Yonsei University, Seoul, Korea in 2003 and 2005, respectively. Currently, he is a researcher at Yonsei University, Seoul, Korea, studying on various human-centered information-processing projects. He also runs a consulting company that provides signal-processing and machine-learning related solutions. Before, he had been with Idiap Research Institute, Switzerland (postdoctoral researcher, Jun. 2011 - Jan. 2013) and University of Southern California, Los Angeles, California (postdoctoral researcher, Jan. 2011 - Jun. 2011).

Maurizio Filippone Maurizio Filippone received a Master's degree in Physics and a Ph.D. in Computer Science from the University of Genova, Italy, in 2004 and 2008, respectively. In 2007, he was a Research Scholar with George Mason University, Fairfax, VA. From 2008 to 2011, he was a Research Associate with the University of Sheffield, U.K. (2008-2009), with the University of Glasgow, U.K. (2010), and with University College London, U.K (2011). He is currently a Lecturer with the University of Glasgow, U.K. His current research interests include statistical methods for pattern recognition. Dr Filippone serves as an Associate Editor for Pattern Recognition and the IEEE Transactions on Neural Networks and Learning Systems.

Fabio Valente (M'06) received the M.Sc. degree (summa cum laude) in communication systems from Politecnico di Torino, Turin, Italy, in 2001, the M.Sc. degree in image processing from University of Nice-Sophia Antipolis, Nice, France, in 2002, and the Ph.D. degree in signal processing from University of Nice-Sophia Antipolis for his work on variational Bayesian methods for speaker diarization done at the Institut Eurecom, France, in 2005. In 2001, he was with the Motorola Human Interface Lab (HIL), Palo Alto, CA. Since 2006, he has been with the Idiap Research Institute, Martigny, Switzerland, and is involved in several EU and U.S. projects on speech and audio processing. His main interests are in machine learning and speech recognition.

Alessandro Vinciarelli (M'06) is Senior Lecturer at the University of Glasgow (UK) and Senior Researcher at the Idiap Research Institute (Switzerland). His main research interest is in Social Signal Processing, the domain aimed at modelling analysis and synthesis of nonverbal behaviour in social interactions (`www.dcs.gla.ac.uk/vincia`). Overall, Alessandro has published more than 90 works, including one authored book, three edited volumes, and 25 journal papers. Furthermore, he is or has been Principal Investigator of several national and international projects, including a European Network of Excellence (the SSPNet, `www.sspnet.eu`), an Indo-Swiss Joint Research Project and an individual project in the framework of the Swiss National Centre of Competence in Research IM2 (`www.im2.ch`). Last, but not least, Alessandro is co-founder of Klewel (`www.klewel.com`), a knowledge management company recognized with several awards.